특집 05

# 인공지능 시스템의 신뢰성 확보를 위한 설명가능 인공지능 기술 동향

김근형 (동의대학교)

목 차



## 1. 서 론

최근 인공지능(AI: Artificial Intelligence) 기술은 많은 애플리케이션의 필수적인 부분이 되었다. 많은 분야에서 인공지능 기반 의사결정 알고리즘의 확산이 급속히 이루어진 이유는 GPU 컴퓨팅 및 클라우드 컴퓨팅과 같은 강력한 컴퓨팅 자원을 활용하여 다양하고 방대한 데이터 처리가 가능해지면서 새롭고 강력한 알고리즘의 개발이 가능해졌기 때문이다[1]. 그러나 많은 애플리케이션에서 성공적으로 활용되고 있는 인공지능 모델인 심층 신경망(DNN: Deep Neural Network)은 모델이나 알고리즘이 어떻게 작동하고 의사결정을 도출하는지 충분한 단서를 제공하지 않아 블랙박스로 인식되었다. 인공지능 알고리즘의 의사결정은 인간의 이익, 인권, 삶에 영향을 줄 수 있으며, 특히 금융의 신용 승인, 사이버 보안의 침입탐지, 국방의 자동화기 등과 같은 임무 수행에 필수적인 기능을 갖는 애플리케이션의 경우 영향이 매우 크기 때문에 결정의 이유를 설명하는 것은 신뢰성 관점에서 매우 중요하다. 규제기관은 인공지능의 자동화된 의사 결정 시스템의 공정성, 책임성 및 투명성 등을 중점적으로 다루기 위해서 “설명할 권리(right to explanation)”라 불리는 유럽연합의 일반 데이터보호규정[2](GDPR: General Data Protection Regulation)[2], 미국 정부의 “2019년 알고리즘 책임법안[3](Algorithmic Accountability Act of 2019)” 그리고 미국 국방부의 인공지능 윤리원칙[4]과 같은 새로운 법률을 제정하였다.

인공지능의 확산이 활발하게 이루어지면서 부작용으로 다양한 형태의 문제가 발생하고 있다. 부작용의 예로 편견, 편향, 차별, 윤리 문제 등이 있다. 인공지능은 인공지능이 분석하는 사회의 구조와 역학 관계에서 학습이 이루어지기 때문에 현실 데이터를 이용하는 인공지능 알고리즘은 사회의 소외, 편견, 불평등, 차별의 패턴을 확대 증폭시킬 수 있다. 또한 데이터마이닝을 위한 모델의 특징, 측정 기준 및 분석 구조 중 많은 부분에 설계자

가 관여하기 때문에 설계자의 선입견과 편견을 복제할 수도 있다. 언어를 훈련하는 인공지능 모델은 사람의 편견을 함께 배우게 되며, 편향된 학습 데이터를 사용하여 학습한 인공지능 모델은 편향된 결정을 할 위험이 크다. 아마존에서 만든 인공지능 채용 시스템은 대부분 10년 이상 경력의 남성 지원자를 추천하였으며 구글의 음성인식 시스템은 특정 지역의 음성 및 남성 음성을 정확하게 인식하는 차별을 보였다[5].

인공지능 기술 활용 이전에는 인지 기능이 책임 있는 인적 대리인에게만 국한되었는데 사람이 인공지능 시스템의 결정, 예측 또는 분류에 영향을 받는 경우 결과에 대한 책임을 물을 수 있는 당사자를 특정하지 못하는 문제가 있다. 특히 부정적인 결과가 발생하는 경우 영향을 받는 개인은 자신의 자율성과 권리를 침해받을 수 있다. 무책임한 데이터 관리, 부주의한 인공지능 시스템 설계 및 추론 프로세스, 의심스러운 네트워크 구성 각각은 신뢰할 수 없는 결과를 추론하는 인공지능 시스템의 구현과 배포로 이어질 수 있으며 결국 정확성이 낮은 결정은 유익한 인공지능 기술의 책임 있는 사용에 대한 신뢰를 해칠 수 있다.

앞서 언급한 부작용 때문에 2017년에 DARPA (Defense Advanced Research Projects Agency)는 설명가능 인공지능 프로그램을 시작하였다. 이 프로그램은 머신러닝 기술에 효과적인 설명 기법을 결합하여 해당 모델의 관련자가 인공지능 시스템을 이해하고 적절하게 신뢰하며 효과적으로 관리할 수 있는 설명가능 인공지능 모델을 생성하는 새로운 또는 수정된 머신러닝 기술 개발을 목표로 한다. 인공지능 모델이 어떤 근거로 의사결정을 내렸는지 알 수 있도록 설명 가능성을 추가하는 것은 사용자 또는 머신러닝과 연관된 사람이 인공지능 모델을 신뢰할 수 있도록 한다. 또한 인공지능 알고리즘 설계자와 의사 결정자를 합리적으로 설득할 수 있으며 인공지능 모델이 어떠한 상황에서 동작하는지 또는 동작하지 않는지, 인공지능 시스템이 왜 실패하고 성공하는지를 파악할 수 있어 설명 가능성은 인공지능 개발자 입장에는 인공지능 알고리즘을 개선하기 위해 사용할 수 있고 비전문가 입장에는 인공지능 알고리즘을 신뢰할 수 있는 근거가 될 수 있다.

구체적으로 합성곱신경망(CNN: Convolution Neural Network)과 순환신경망(RNN: Recurrent Neural Network)과 같은 심층 신경망은 이미지/비디오 처리 및 컴퓨터 비전 작업에서 높은 성능을 보여 최근 많은 영역에서 활용되고 있다. 그러나 이러한 심층 구조는 비선형적인 요소가 많아 완전히 투명하지 않고 매우 복잡하여 임무 수행에 필수적인 애플리케이션에 이러한 복잡한 모델을 적용하기 전에 입력 데이터 또는 신경망 내의 기능이 어떠한 의사결정에 영향을 주는지 이해하는 것이 중요하다. 따라서 예측한 결정을 검증하여 심층 모델에 대한 신뢰 관계를 구축하고, 보이지 않거나 익숙하지 않은 실제 데이터에 대해서 기대한 것처럼 안정적으로 작동하는지 확인해야 한다. 특히 생/의학 애플리케이션, 자율 주행과 같은 핵심 분야에서는 한 번의 잘못된 결정이 치명적인 결과를 초래하고 인명까지 위협할 수 있어 한 번의 잘못된 결정도 용납할 수 없다. 인공지능 모델 내 알고리즘의 모델 정확성에 상관없이 결정에 대한 신뢰성을 보장하기 위해서는 결정에 대한 합리적인 이유를 이해하고, 분석하고, 시각화하고, 한마디로 설명하는 것이 필수적이다.

본 고에서는 인공지능 시스템의 신뢰성 제고를 위해 논의되고 있는 설명가능 인공지능 개념, 원칙, 설명의 범주 및 설명기술 분류를 살펴보고 DARPA의 설명가능 인공지능 프로그램의 주된 연구 분야와 최근 영상 분석과 자연어 처리에 널리 활용되는 심층 신경망에 설명 가능성 지원 기

술 분류를 살펴본다.

## 2. 설명가능 인공지능 개요

### 2.1 설명가능 인공지능 개념 및 연구 추세

설명가능 인공지능의 개념은 1975년 부정확한 의사결정 과정을 조건부 확률 근사 값 방식으로 모델링하는 “설명 가능한 의사결정 체계”에서 처음 등장하였고 1970년대 등장한 의료 진단 전문가 시스템이 도출 결과를 전문가에게 이해시키지 못하면서 설명가능 인공지능의 중요성이 대두되었으며 소수 연구자에 의해서만 연구가 진행되었다. 2004년 반 렌트, 피셔, 만쿠스는 군대의 모의 전투 프로그램 속 NPC (Non-Player Character) 인공지능을 개조하여 행동의 이유를 설명하는 아키텍처를 제안하였고 설명가능 인공지능(XAI: explainable AI) 용어를 처음 사용하였다[6]. 이들은 게임 속 상황을 벡터 형태로 정리하는 제어 인공지능과 NPC 객체 행동의 결과를 분석해서 명령을 내리는 역할을 하는 명령 인공지능으로 구성하여 명령 인공지능의 상태 인지 결과와 제어 인공지능의 지시를 분석하여 행동의 이유를 이해하였다. 2016년 심층 신경망 기술을 활용한 알파고가 이세돌과의 대국에서 프로기사들이 이해하지 못하는 위치에 바둑돌을 두었는데, 이는 알파고의 승리를 결정짓는 결정적인 한 수가 되었다. 이 경우 알파고의 블랙박스 특성은 문제가 되지 않았으나 많은 애플리케이션에서 인공지능 모델의 의사결정 과정을 이해하고 검증하는 것이 불가능하다는 것은 분명 인공지능의 단점이고 인공지능 활용의 장애가 된다. 그러므로 설명가능 인공지능에 관한 관심이 (그림 1)에 보인 것처럼 2017년 이후 급격하게 증가하고 있다. 또한, ArXiv에 공개된 컴퓨터 공학 분야 논문 중 제목에 “Explainable AI” 키워드를 포함한 논문이 2015년 9건에서 2019년 236건으로 증가하였고 매년 전년 대비 2배 이상의 급격한 증가를 하였다[7].

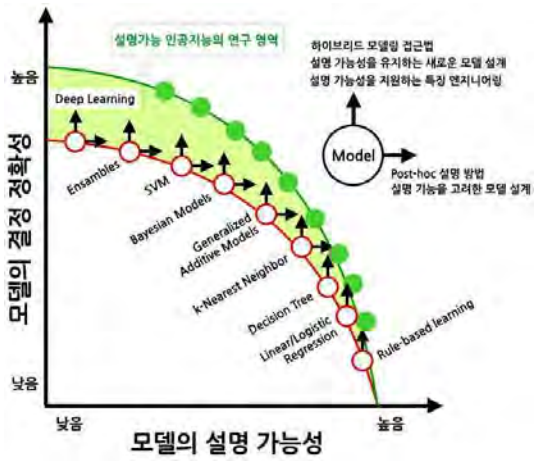
많은 연구가 해석 가능성과 설명 가능성을 위해 다양한 방안과 프레임워크를 제시하고 있으며 설명 가능한 인공지능은 머신러닝 연구 커뮤니티로부터 많은 관심을 받고 있다. 설명가능 인공지능은 원칙/법을 옹호하고 설명가능 의사결정 시스템 및 연구를 촉진할 필요성이 계속 증가함에 따라 다시 떠오르는 연구 주제가 되었다. 최근 인공지능과 머신러닝의 발전과 광범위한 영역으로 적용 확대, 비윤리적인 사용, 투명성 부족, 모델의 바람직하지 못한 편향에 대한 높은 우려 때문에 인공지능 시스템의 신뢰성 제고를 위한 설명가능 인공지능 연구에 관한 관심이 높아졌다.

### 2.2 인공지능 알고리즘 정확성과 설명 가능성

산업계에서는 규칙 기반 학습, 선형 모델(선형/


(그림 1) “Explainable AI”로 검색한 Google Trends Popularity Index (최대값: 100).

로지스틱 회귀분석), 결정트리 모델과 같이 직관적이며 검증이 쉽고 좀 더 해석이 가능한 모델을 선호하는 경향이 있다. 이와 대조적으로 실제 데이터의 복잡한 구조를 고려하면 모델 선택 및 구축 단계에서 대부분 관심이 심층 신경망과 같은 보다 복잡한 모델로 집중된다. 복잡한 모델을 사용하면 개선된 결과를 얻을 수 있다. 그러나 이와 같은 모델(앙상블 모델, 딥 러닝 등)은 블랙박스 모델이라 여겨지고 있다. 모델이 발전하여 결과의 정확성이 높아질수록 작동 방식을 설명하기가 더 어려워진다. (그림 2)는 DARPA에서 분석한 다양한 머신러닝 모델 결과의 정확성과 설명 가능성 간의 균형 관계를 보인다. 머신러닝 모델의 알고리즘이 복잡할수록 설명 가능성이 작고 머신러닝 모델의 알고리즘이 단순할수록 설명 가능성이 크다. 최근 많은 영역에서 사용되는 심층 신경망은 모델이 복잡하여 설명 가능성이 작으나 정확성은 높아 많이 사용되고 있다. (그림 2)의 설명가능 인공지능의 연구 영역으로 표시된 영역은 앞으로 각 인공지능 모델의 정확성과 설명 가능성 관점에서 개선할 수 있는 정도를 보인다. 각 모델에서 설명




(그림 2) 머신러닝 모델의 정확성과 설명 가능성 균형 관계 [9].

가능성을 높이기 위해 모델의 결과를 바탕으로 결과의 근거를 찾는 방법(post-hoc 방법) 또는 모델을 설계할 때부터 해석 기능을 고려하는 방법이 고려되고 있다. 모델 결과의 정확성을 높이기 위해서 하이브리드 모델링, 설명 가능성을 유지하는 새로운 모델 설계 또는 설명 가능성을 지원할 수 있도록 특징을 엔지니어링 하는 방법이 고려된다.

## 2.3 설명가능 인공지능 시스템 원칙

설명가능 인공지능 시스템은 먼저 설명을 제공하여야 하며, 설명을 사용하려는 대상에게 의미가 있어야 한다. 그리고 설명의 정확성이 높아야 하며 시스템이 처리할 수 있는 지식의 한계를 제시하여야 한다는 원칙을 만족하여야 한다[8]. 각 원칙에 대한 정의 및 상세 설명은 다음과 같다.

- 설명(Explanation) 제공: 설명가능 인공지능 시스템은 특정 결과에 대해 수반되는 증거 또는 이유를 제공하여야 한다. 설명 제공 원칙은 인공지능 시스템이 각 결과에 대한 증거와 추론을 제공하도록 한다. 이 원칙은 설명가능 인공지능 시스템이 설명을 제공하여야 한다고 명시할 뿐 설명의 정확성, 유용성, 이해 가능성을 요구하지 않으며 설명에 대한 품질 측정 기준도 명시하지 않는다.
- 의미 있는(Meaningful) 설명 제공: 설명가능 인공지능 시스템은 개별 사용자가 이해할 수 있는 설명을 제공하여야 한다. 일반적으로 개인이 설명을 이해할 수 있거나 설명이 작업 수행에 유용하면 이 원칙이 충족되었다고 볼 수 있다. 그러나 이 원칙은 하나의 설명이 모든 경우에 적합하다는 것을 의미하는 것은 아니며 하나의 시스템에 대해 여러 그룹의 사용자가 다른 설명을 요구할 수 있다. 의미 있는 설명 원칙은 각 사용자 그룹에 맞춰진 설명을 요구하며 각 사용자

그룹의 목표와 필요조건이 다를 수 있다. 예를 들어 설명가능 인공지능 시스템의 설명 중 법의 학자에게 의미가 있는 것과 배심원에게 의미가 있는 것은 다를 수 있다. 두 사용자 그룹은 인공지능 시스템의 결정을 다양한 이유로 다른 방식으로 해석할 수 있다. 한 가지 이유는 개인의 사전 지식과 경험이 결정에 영향을 주기 때문이고 다른 이유는 사람들 사이의 심리적 차이가 설명을 해석하는 방법과 어떤 유형의 설명을 의미 있게 생각하는가에 영향을 주기 때문이다. 따라서 사용자마다 같은 인공지능 설명을 다른 의미로 해석할 수 있다. 설명의 의미는 인공지능 시스템의 설명과 개인의 사전 지식, 경험 및 정신적 사고 과정의 조합에 의해 영향을 받는다. 설명의 의미에 영향을 미치는 모든 요소는 인공지능 시스템과 인간 사이의 인터페이스를 구축할 때 어려움을 초래한다. 사람에게 의미 있는 설명을 하는 인공지능 시스템을 개발하려면 알고리즘적 요인과 심리적인 요인 모두를 고려해야 한다.

- 설명 정확성(Explanation Accuracy): 설명은 설명가능 인공지능 시스템의 결과를 생성하는 프로세스를 정확하게 반영하여야 한다. 설명 제공 및 의미 있는 설명 원칙은 설명가능 인공지능 시스템이 사용자 커뮤니티에 의미 있는 설명 제공을 요구한다. 즉, 설명 제공 및 의미 있는 설명 원칙은 설명가능 인공지능 시스템이 결과를 도출하기 위한 내부 프로세스를 정확히 반영하는 설명을 요구하지 않는다. 설명 정확성 원칙은 인공지능 모델의 설명에 정확성을 요구하는 것으로 인공지능 모델 결정의 정확성과 다른 개념이다. 인공지능 모델의 결정 정확성은 모델의 의사결정이 정확한지 아닌지를 표현하고 설명 정확성은 인공지능 모델의 의사결정 정확성과는 관계없이 해당 설명이 인공지능 모델의 결과를 정확하게 설명하는지 아닌지를 뜻한다. 인공지능 모델 결과의 정확성에 대한 측정 지표와 관련한 연구가 많이 이루어졌으나 설명 정확성을 위한 측정 지표의 개발은 초기 단계로 앞으로 연구가 요구된다. 의미 있는 설명 원칙과 마찬가지로 설명 정확성은 그룹과 개인에 대해 서로 다른 측정 기준을 정한다. 일반 사용자는 출력 생성을 위한 프로세스를 완전히 특성화하는 데 필요한 미묘한 차이는 고려하지 않고 간결하게 임계점에 초점을 맞춘 간단한 설명만 필요로 할 것이며 프로세스를 특성화하는 미묘한 차이는 전문가에게만 의미가 있을 수 있다. 전반적으로 한 가지 유형 이상의 설명을 생성할 수 있는 인공지능 시스템은 더 설명 가능한 것으로 여겨질 수 있으나 설명의 수준이 달라 설명 정확성 평가에 사용되는 측정 지표가 보편적이거나 절대적이지 않다.
- 지식 한계(Knowledge Limits): 설명가능 인공지능 시스템은 설계된 조건 범위 내 또는 결과에 대한 충분한 신뢰가 있을 때 동작해야 한다. 앞의 원칙들은 인공지능 시스템이 시스템 설계 시 설정한 지식 범위 내에서 동작한다고 암묵적으로 가정하고 있다. 이 지식 한계 원칙은 시스템이 작동하도록 설계 또는 승인되지 않았거나 해당 답변이 신뢰할 수 없는 사례를 식별하여야 한다는 것이다. 지식 한계를 식별하고 선언하여 인공지능 시스템의 결과가 부적절할 때 판단을 제공하지 않도록 하는 것이다. 지식 한계 원칙은 오해의 소지가 있거나 위험하거나 부당한 결정 또는 결과를 방지하여 인공지능 시스템에 대한 신뢰를 높일 수 있다. 인공지능 시스템이 지식 한계에 도달할 수 있는 경우는 두 가지가 있다. 첫 째는 질문이 시스템의 지식 범위를 벗어나는 경우이다. 예를 들어 조류의 종을 분류하는 인공지능 시스템에 사용자가 사과 이미지

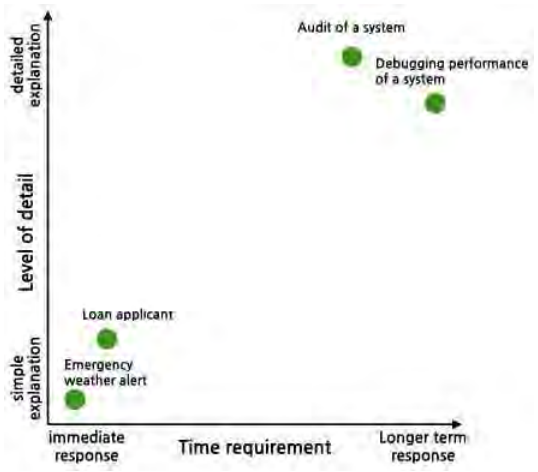
를 입력할 수 있다. 시스템이 입력 이미지에서 새를 찾을 수 없다는 답변을 반환할 수 있어 인공지능 시스템은 결정을 제공할 수 없다. 이것은 답변이며 설명이다. 지식 한계에 도달할 수 있는 두 번째는 가장 가능성이 있은 답변의 신뢰도가 내부 신뢰 임계값에 비해 작은 경우이다. 예를 들어, 조류 분류 시스템의 경우, 새의 입력 이미지가 너무 흐릿하여 종을 결정할 수 없을 때 인공지능 시스템은 이미지가 새인 것으로 인식할 수 있으나 이미지의 품질이 낮다는 것을 인식할 수 있다. 출력의 예로 “이미지에서 새를 찾았으나 이미지 품질이 너무 낮아 식별할 수 없다”가 될 수 있다.

## 2.4 설명의 범주

앞의 원칙에서 살펴본 것처럼 설명은 사용자에 따라 다를 수 있다. 일부 설명은 간단할 수 있지만 다른 설명은 상세하고 완전히 이해하기 위해 교육이나 전문 지식이 요구되는 경우가 있다. [7]에서는 설명의 범주를 논의하기 위해 설명을 다음의 다섯 가지의 범주로 분류하였다.

- 사용자 이익: 이 범주의 설명은 인공지능 시스템의 결과가 사용자 이익과 관련된 것으로 사용자에게 알리기 위한 설명이다. 예로 인공지능 모델이 결정한 대출 신청의 승인 또는 거부된 이유를 신청자에게 설명하여야 한다.
- 사회적 수용: 이 범주의 설명은 사회에서 인공지능 시스템의 신뢰와 수용을 위해 생성되는 설명이다. 예로 인공지능 시스템에서 예측하지 못한 결정이 제공되는 경우 설명은 사용자가 이 결정이 도출된 이유를 이해하는데 도움이 된다. 또한 이론적 근거가 제공될 수 있다면 인공지능 시스템은 더 나은 편안함을 제공할 것이다.
- 규정 및 준수: 이 범주의 설명은 규정, 안전 표준 등의 준수에 대한 감사를 위한 설명이다. 설명의 대상에는 중요하고 상세한 정보가 필요한 사용자(예: 안전 규제기관)와 시스템과 상호작용하는 사용자(예: 개발자)가 포함된다. 예를 들어 자율 주행 자동차 개발자 또는 감사자가 포함될 수 있다.
- 시스템 개발: 이 범주의 설명은 인공지능 알고리즘 또는 시스템의 개발, 개선, 디버깅 및 유지보수를 쉽게 하는 설명이다. 이 범주의 설명 사용자는 개발자, 제품 관리자와 임원이 될 수 있다. 중요하고 상세한 정보가 필요한 사용자와 시스템과 상호작용하는 사용자가 관여한다. 인공지능의 설명 기능을 활용하여 머신러닝 알고리즘을 디버깅하는 개발자가 존재한다.
- 소유자 이익: 이 범주의 설명은 시스템 운영자에게 도움이 되는 설명이다. 예를 들어 볼 영화나 비디오를 추천할 때 이전에 본 콘텐츠를 기반으로 추천한 선택 항목을 설명하는 추천 시스템이 있다. 추천 시스템은 설명과 함께 영화를 추천했을 때 사용자가 설명을 신뢰하면 사용자가 추천 영화를 시청하기 때문에 소유자에게 이익이 된다.

실용적인 관점에서 설명은 설명에 응답하는 시점과 설명의 상세함에 의해 특징 지울 수 있다. 다음 (그림 3)은 설명의 상세함과 설명에 응답하는 시점 간의 관계를 보여준다. 예를 들어 긴급 기상 경보의 경우 사용자가 긴급히 행동해야 하므로 경보와 간단한 설명이면 충분하다. 디버깅에 대한 설명은 시스템의 내부 단계에 대한 상세한 정보가 포함되어야 하며 사용자가 설명을 검토하고 다음 조치 결정에 시간이 걸릴 수 있다. 대출 신청에 대한 사용자의 대응은 일반적으로 빠르고 설명은 기상 경보보다 더 자세한 정보를 제공한다. 시스템 감사에 대한 사용자의 응답 시간 및 설명




(그림 3) 특정 상황별 설명의 상세함과 응답 시간의 관계 [7].

의 상세함은 시스템 디버깅과 유사할 수 있다.

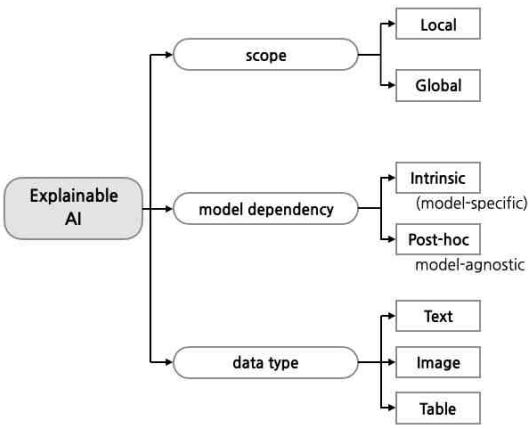
## 2.5 설명가능 인공지능 기술 분류

설명가능 인공지능 기술은 설명 기능이 인공지능 모델에 내재(intrinsic 설명)되어 있는 경우와 인공지능 모델 학습 후 결과를 분석하여 설명을 도출하는 방법(post-hoc 설명)으로 분류한다 [8]. (그림 2)에서 표시한 규칙 기반 학습, 결정트리 또는 희소 선형 모델은 단순한 모델로써 모델의 구조로 모델 결정의 설명이 가능하다. 이처럼 구조가 단순한 모델은 그 자체적으로 이미 설명 가능성을 확보하고 있어 내재적 방법으로 분류된다. 반면 심층 신경망 기반 대부분의 머신러닝 모델을 설명하기 위해서는 머신러닝의 결정을 기반으로 이유를 설명하는 방법(post-hoc 방법)이 주로 적용된다.

또한 설명하는 범위에 따라 설명가능 인공지능 기술을 분류할 수 있다. 인공지능 모델의 모든 결정에 관해서 설명이 가능한 전역적인(global) 방법과 일부 예측 결과만 설명이 가능한 국소적인(local) 방법이 있다. 전역적인 방법은 모델 내부 로직에 관련한 이해를 바탕으로 모든 결정을 설명한다. 내재적 방법은 모델의 구조로부터 모든 예측 또는 결정에 대한 설명이 가능하여 전역적인 방법으로 분류된다. post-hoc 방법으로 모든 예측 또는 결정을 설명하는 경우 모든 개별 결정을 구체적으로 설명하는 것이 다소 어려울 수 있다. 국소적인 방법은 설명할 범위가 좁아 설명 가능성이 크고 비용이 적게 요구된다. 현실적으로 인공지능 모델이 결정을 도출할 때 마다 설명을 할 필요는 없고 설명이 필요할 때 해당 이슈에 대해서 잘 설명하는 것이 더 바람직할 수 있다.

또한 설명 제공 방법은 특정한 인공지능 모델에만 적용할 수 있는 경우와 인공지능 모델과 관계없이 임의의 인공지능 모델에 적용할 수 있는가에 따라 모델 특화된(model-specific) 설명 방법과 모델에 관계없는(model-agnostic) 설명 방법으로 분류된다. 모델과 관계없는 설명 방법은 모델의 내부구조를 알 수 없거나 모델의 특징을 하나도 이용하지 않고 모델 결정의 이유를 모델 외부에서 도출하게 된다. 모델 특화된 설명 방법은 모델 자체가 가지고 있는 특징을 이용하여 타 모델에는 적용할 수 없는 것이다. 내재적 방법과 합성곱신경망 기반의 모델에 사용하는 히트맵 같은 시각화 해석 방법은 모델 특화된 설명 방법에 속한다.

앞에서 살펴본 바와 같이 설명가능 인공지능 기술의 분류는 다양한 관점이 존재한다. 설명가능 인공지능 기술은 구별되고 더 나눌 수 있는 서로 다른 관점이 있어 기술 분류가 일방적이지는 않다. 그러므로 실무자가 직면한 각 문제의 특정한 기준에 대한 이상적인 방법을 선택하기 위해서는 각 기술의 모든 측면을 고려해야 한다. 다양한 설명가능성 모델은 다음 (그림 4)와 같이 구분될 수 있다.



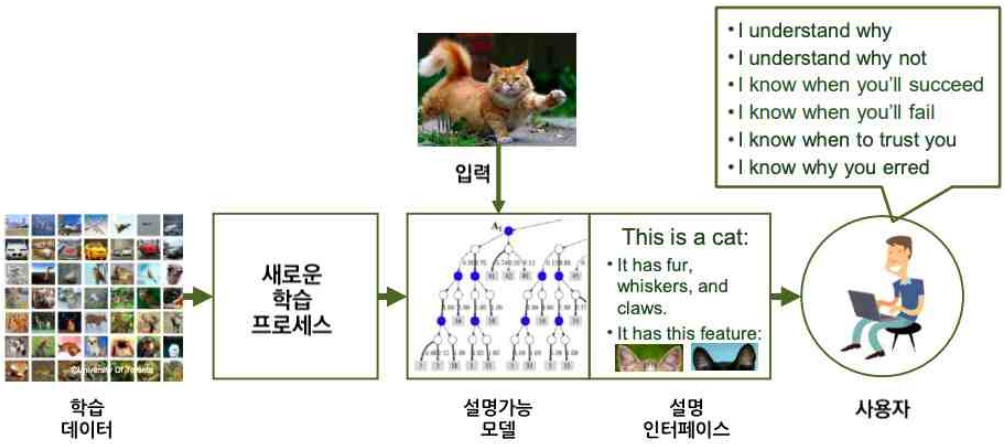

(그림 4) 설명가능 인공지능의 분류 체계.

## 3. 설명가능 인공지능 기술 동향

### 3.1 DARPA의 설명가능 인공지능 기술

DARPA에서 제시한 사용자 중심의 설명가능 인공지능 구조를 (그림 5)에 보인다 [9]. 사용자가 인공지능 모델의 전반적인 강점과 약점을 이해하고 향후 또는 다른 상황에서 어떻게 작동할 것인지 이해하며 사용자가 시스템의 오동작을 수정할 수 있도록 설명을 제공한다. 사용자 중심의 설명가능 인공지능 개념은 더 많은 설명이 가능한 모델을 생산하는 방법, 설명 인터페이스를 설계하는 방법 및 효과적인 설명을 위한 심리적 요건을 이해하는 방법들의 상호 관련된 연구를 수행하였다.

DARPA의 설명가능 인공지능 프로그램은 (그림 5)에 나타낸 설명가능 모델, 사용자를 위한 설명 인터페이스를 중점적으로 다루며 효과적인 설명을 위한 심리학적 요구사항에 대해서도 함께 검토하고 있다. DARPA의 설명가능 인공지능 모델은 머신러닝 기술을 수정하거나 새로운 머신러닝 기술 개발을 통해 모델의 학습 정확도를 높게 유지하면서 설명 가능성을 높이는 연구를 수행하며 심층 설명(deep explanation), 해석가능 모델(interpretable model), 모델 귀납(model induction)의 세 가지 전략을 통해 개발한다 [8]. 세 전략은 적용해야 할 상황과 효용에 따라 선별적으로 사용할 수 있다. 심층 설명 전략은 설명할 수 있는 특성을 학습하기 위해 심층 신경망을 수정하는 것이다. 이를 위한 기술 중 하나가 신경망의 여러 계층에 요소를 추가하여 학습 중에 이루어지는 복잡한 연결을 이해하는 방법이다. 해석가능 모델은 더욱 구조화되고 해석 가능한 인과 모델을 학습하는 머




(그림 5) DARPA의 설명가능 인공지능 구조 [8].

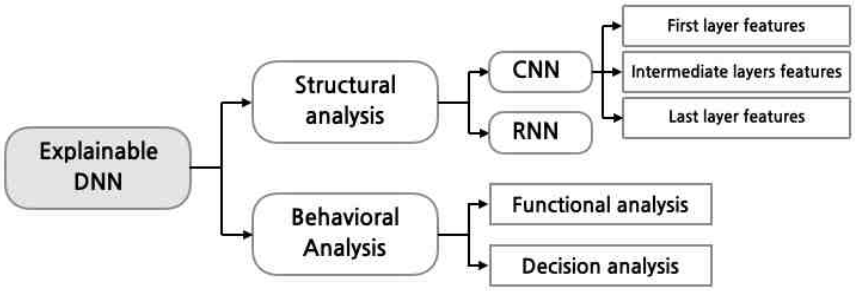
신러닝 기술이다. 해석가능 모델의 예로 BPL (bayesian program learning) 모델과 확률적 AND-OR 그래프 기반 모델이 있다. 모델 귀납 전략은 모든 머신러닝 모델을 블랙박스로 간주하고 실험하여 대략적인 설명가능 모델을 추론하는 기술로 앞에서 분류한 Post-hoc 설명 방법에 속한다. 모델 귀납의 예로 로컬 대리 분석 모델(Local Surrogate Analysis Model) 중 하나인 LIME (Local Interpretable Model-agnostic Explanations) 모델이 있으며 인공지능 모델에 관계없이 적용할 수 있는 방법이다. DARPA의 설명 인터페이스 개발은 HCI(Human Computer Interaction) 기술을 이용하여 인공지능 모델의 의사결정에 대한 설명을 사용자가 이해할 수 있는 방식으로 표현하는 연구이다.

DARPA의 설명가능 인공지능 프로그램은 인공지능 운영과 관련하여 데이터 분석(지도 학습)과 자율성(강화 학습)의 두 가지 문제 영역을 다룬다. 데이터 분석 문제는 이종 멀티미디어 데이터에서 관심이 있는 객체의 분류에 관련한 것으로 지식 분석가는 빅 데이터 분석 알고리즘의 의사결정 및 추천 사항을 시스템으로부터 제시받으며 분석에서 어떤 내용을 결정의 증거로 보고할 것인지 그리고 어떤 것으로 더 추구할 것인지 결정한다. 자율성 문제는 강화 학습 과제와 관련이 있고 시뮬레이션된 목표에 대한 의사결정 정책을 학습하는 것으로 인공지능 파트너를 효과적으로 관리해야 하는 필요성 때문에 비롯되었다. 예를 들어 국방부가 전투원 능력을 높이기 위해 반자동 시스템을 개발한다면 향후 임무에서 이러한 기능을 가장 잘 사용할 방법과 시기를 결정하기 위해서 기능이 어떻게 동작하는지 이해해야 한다. 이 경우에도 효과적인 설명은 전투원이 올바른 결정을 할 수 있도록 할 것이다. DARPA는 두 도전 영역 모두에서 설명의 정확성 또는 효율성의 측정을 중요한 주제로 설정하였다. 학습된 모델의 설명 정확성을 자동으로 측정할 수 있다면 편리하겠지만 설명가능 인공지능 시스템이 사용자에게 도움이 되는지를 평가해야 한다. 이를 위해 사용자의 만족도, 멘탈 모델, 작업의 성과 및 적절한 신뢰를 측정하기 위한 심리학 실험 및 지식 한계의 범위를 설정해야 한다.

## 3.2 심층 신경망의 설명 가능 인공지능 기술

심층 신경망의 중요성에도 불구하고 심층 신경망의 설명 가능성에 관한 연구는 아직 초기 단계에 있다. 심층 신경망의 설명가능 기술은 (그림 4)의 데이터 유형에 따라 시각적, 텍스트 기반 또는 예제 기반으로 구현되고 있다 [10]. 일반적으로 설명가능 심층 신경망 모델의 주요 목표는 (1) 심층 신경망 모델 내부에서 무슨 일이 일어나고 있나? (2) 심층 신경망 모델의 각 계층은 어떤 역할을 하나? (3) 심층 신경망 모델은 어떤 특징을 찾나? (4) 왜 심층 신경망 모델의 결정을 믿어야 하나? 와 같은 질문에 대해 답을 하는 것이다. (그림 6)은 설명가능 심층 신경망의 분석 기술을 post-hoc 설명 방법으로 구조 분석과 동작 분석으로 분류하고 구조 분석은 심층 신경망을 합성곱신경망과 순환신경망의 내부구조 분석을 위해 사전 훈련된 신경망의 특성을 시각화하며 동작 분석은 신경망의 전반적인 동작과 내부 동작의 설명에 해당한다. 고정된 구조로 사전 학습된 신경망을 설명하는 대신 설명 기능을 신경망 설계 시 고려하는 설명 가능성을 포함한 모델 설계 방법이 있다. 각 방법의 설명은 다음과 같다.

- 구조 분석: 합성곱신경망의 학습된 커널 및 순환신경망의 데이터 구조를 이해하기 위해 먼저 사전 학습된 네트워크의 기능을 시각화하는 심층 신경망 시각화 및 이해로 부터 시작한다. 이러한




(그림 6) 심층 신경망의 설명 방법 분류 [10].

관점에서 심층 신경망의 동작을 분석한다. 합성곱신경망은 다시 첫 번째 계층, 중간 계층, 마지막 계층으로 나누어 각 계층의 특징을 분석한다.

- 동작 분석: 심층 신경망의 전반적인 동작과 내부 동작을 설명 가능성 개념으로 구체적으로 설명한다. 전반적인 동작(기능 분석)은 입출력 관계에 기반하고 내부 동작(의사결정 분석)은 결정을 설명하기 위해 수학적인 분석을 한다. 기능 분석은 전체 신경망을 블랙박스로 간주하고 입력 이미지에 대한 특정 결정에 가장 적합한 픽셀을 찾음으로써 설명 가능성을 제공한다. 즉 네트워크의 작동을 입력과 출력 간의 관계를 발견하여 해석한다. 의사결정 분석은 의사결정에 있어 어떤 뉴런이 더 중요한 역할을 하는지 보여주기 위해서 신경망을 통해 결정을 분해하고 역 추적하여 신경망의 결정을 설명하려는 방법으로 최종 결정과 관련성은 입력 데이터의 기여를 시각화하는 히트맵을 사용한다.
- 설명 가능성을 포함한 모델 설계: 심층 신경망을 구축한 후 설명을 위한 지식을 구축하는 것은 설명 가능성을 염두에 두고 구조와 알고리즘을 개발하지 않으면 설명이 어렵고 비용이 많이 든다. 그래서 내재된 설명 방법으로 설명 가능성을 포함하는 심층 신경망 설계에 관한 연구가 진행 중이다.

## 4. 결 론

설명가능 인공지능 기술은 향후 보편화될 인공지능 모델의 신뢰성을 높이기 위해서 필수적인 부분이다. 본고에서는 설명가능 인공지능의 개념, 원칙 및 설명의 범주, 분류 체계를 살펴보았으며 설명 모델, 설명 인터페이스로 구성된 DARPA의 설명가능 인공지능 구조와 도전 과제로 설정한 데이터 분석 문제와 자율성 문제 그리고 설명 가능 심층 신경망에 관한 연구 동향을 살펴보았다.

설명가능 인공지능 시스템의 설명은 업무와 사용자에 따라 요구사항과 설명의 필요성이 다르다. 설명이 필요한 시스템의 범위를 다룰 때 설명의 상세함과 응답 시간을 고려하여야 한다. 또한 지식 한계를 이해하고 입력 데이터가 인공지능 시스템의 범위를 벗어남을 선언하는 알고리즘 개발이 필요하다. 또한 설명 정확성도 중요한 요소로 이를 측정하는 연구도 매우 제한적으로 이루어지고 있어 관련 연구가 필요하다.

### 참 고 문 헌

[1] S. R. Islam, W. Eberle, S. Khaled, M. Ahmed, “Explainable Artificial Intelligence Approaches: A Survey,” arXiv.2101.09429

v1, Jan. 2021.
[2] KISA GDPR 대응지원 센터, https://gdpr.kisa.or.kr.
[3] 신용우, “인공지능 관련 입법 현황 및 전망,” 국회입법조사처, 2019년 12월
[4] Depart Innovation Board, “AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense,” Oct. 2019.
[5] 양서연, “차별하는 인공지능: AI가 확산할 수 있는 은연 중 차별과 우려와 새로운 사회적 윤리의 필요성,” pp. 28-34, 한국여성정책연구원 젠더리뷰 54호, 2019.
[6] M. V. Lent, W. Fisher, M. Mancuso, “An Explainable Artificial Intelligence System for Small-unit Tactical Behavior,” pp. 25-29, 16th Proc. of Innovative Applications of Artificial Intelligence, July, 2004.
[7] P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, M. A. Przybocki, “Four Principles of Explainable Artificial Intelligence,” NIST Report, Draft NISTIR 8312, August 2020.
[8] A.B. Arrieta, N. Diaz-Rodriguez, J. D. Ser, A. Bennetot, S. Tabik, et. al., “ Explainable Artificial Intelligence(XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsive AI,” Information Fusion, Vol. 58, Dec. 2019.
[9] D. Gunning, D. W. Aha, “DARPA’s Explainable Artificial Intelligence Program,” AI Magazine, Vol. 40, No. 2, pp. 44-58, Summer 2019.
[10] W. Samek, T. Wiegand, K.-R. Muller, “Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models,” ITU Journal: ICT Discoveries, Special Issue No. 1, Oct. 2017.

## 저 자 약 력

**김 근 형**

이메일 : geunkim@deu.ac.kr

- 1986년 서강대학교 전자공학과 (학사)
- 1988년 서강대학교 전자공학과 (석사)
- 2005년 포항공과대학교 컴퓨터공학과 (박사)
- 1988년~1990년 (주)LS산전 / 연구원
- 1990년~1993년 삼성종합기술원 / 선임연구원
- 1993년~2007년 (주)KT / 수석연구원
- 2007년~현재 동의대학교 게임공학 전공 교수
- 관심 분야 : 설명가능 인공지능, Web ML(Machine Learning), 딥러닝